

Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles

So-Jung Park · Irina Kufareva · Ruben Abagyan

Received: 8 December 2009 / Accepted: 26 April 2010 / Published online: 9 May 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract Nuclear receptors (NRs) are ligand dependent transcriptional factors and play a key role in reproduction, development, and homeostasis of organism. NRs are potential targets for treatment of cancer and other diseases such as inflammatory diseases, and diabetes. In this study, we present a comprehensive library of pocket conformational ensembles of thirteen human nuclear receptors (NRs), and test the ability of these ensembles to recognize their ligands in virtual screening, as well as predict their binding geometry, functional type, and relative binding affinity. 157 known NR modulators and 66 structures were used as a benchmark. Our pocket ensemble library correctly predicted the ligand binding poses in 94% of the cases. The models were also highly selective for the active ligands in virtual screening, with the areas under the ROC curves ranging from 82 to a remarkable 99%. Using the computationally determined receptor-specific binding energy offsets, we showed that the ensembles can be used for predicting selectivity profiles of NR ligands. Our results

evaluate and demonstrate the advantages of using receptor ensembles for compound docking, screening, and profiling.

Keywords Nuclear receptor · Multiple receptor conformers · Virtual ligand screening · Ligands profiling · Internal Coordinate Mechanics

Introduction

The primary function of nuclear receptors (NR) is to act as transcriptional switches responding to lipophilic hormones, vitamins, dietary lipids, or other intracellular signals [1, 2]. The common modular structure of an NR includes a conserved N-terminal DNA-binding domain (DBD), and a C-terminal ligand-binding domain (LBD) [3]. The DBD directly interacts with DNA response element on the target genes to modulate transcription. The role of the LBD is to recognize ligands and coregulators [4]. Ligand binding to NRs LBD induces changes in the conformational equilibrium and dynamic behavior of receptors and regulates the recruitment of coregulators and chromatin-modifying machineries [5, 6].

Nuclear receptors can be classified into the three groups. The first is the classic and the most extensively characterized group of receptors that respond to steroid hormones, thyroid hormone (thyroid hormone receptor; TR), and vitamins A (retinoic acid receptors; RAR) and D (vitamin D receptor; VDR). These receptors are all essential for homeostatic control of endocrine system. The second class is the orphan nuclear receptors, which are structurally related to nuclear hormone receptors but lack known natural ligands. In this group, nerve growth factor-induced clone B (Nur77), nuclear receptor related-1 (Nurr1), and neuron-derived orphan receptor-1 (NOR1) are involved in the inflammatory processes [7]. The third class of nuclear

Electronic supplementary material The online version of this article (doi:[10.1007/s10822-010-9362-4](https://doi.org/10.1007/s10822-010-9362-4)) contains supplementary material, which is available to authorized users.

S.-J. Park
Department of Molecular Biology, The Scripps Research
Institute, 10550 N Torrey Pines Rd., La Jolla, CA 92037, USA

I. Kufareva · R. Abagyan (✉)
Skaggs School of Pharmacy and Pharmaceutical Sciences,
9500 Gilman Drive #0747, La Jolla, CA 92093-0747, USA
e-mail: rabagyan@uscd.edu

R. Abagyan
Molsoft, LLC, 3366 N Torrey Pines Ct. Suite 300,
La Jolla, CA 92037, USA

receptors includes adopted orphan nuclear receptors, which function as heterodimers with retinoid X receptor (RXR). These include receptors for fatty acids (peroxisome proliferator activated receptors; PPAR α , δ , γ), oxysterols (liver X receptors: LXR α , β), bile acids (farnesoid X receptor; FXR), xenobiotics (steroid and xenobiotic receptor/pregnane X receptor; SXR/PXR [8]) and constitutive androstane receptor; CAR) [9].

NRs are transcription factors naturally switched on and off by small-molecule hormones. These hormones often present physicochemical properties very similar to therapeutic chemical entities. As such, NRs represent an important class of drug targets that have potential applications in cancer, inflammatory diseases, and diabetes.

Several NRs are already targeted by known drugs, such as tamoxifen, a partial antagonist of estrogen receptor (ER), or bicalutamide, an antagonist of androgen receptor (AR) used in treatment of breast and prostate cancer, respectively [10, 11]. Progesterone receptor (PR) agonists have several important applications in women's health, such as in oral contraception and post-menopausal hormone replacement therapy [12]. Spirolactones are mineralocorticoid receptor (MR) antagonists that have been used in the treatment of sodium-retaining states and as antihypertensive agents [13]. The therapeutic applications of vitamin D metabolites are treatments for renal osteodystrophy, osteoporosis, psoriasis, cancer, autoimmune disease, or prevention of graft rejection [14].

The rapid growth of the number of high quality crystal structures of ligand binding domains of a number of important nuclear receptors [15–18] creates an opportunity to build a foundation for structure-based identification of potential nuclear receptor agonists, antagonists, or tissue selective modulators [19]. The incorporation of multiple conformers of the same ligand binding pocket, a.k.a. ensemble docking [20] results in a higher quality algorithms for ligand binding pose prediction, virtual compound screening, and selectivity profiling. The recently developed 4D docking technology [21] further improves this process by dramatically speeding up all computational steps.

We have previously demonstrated the use of multiple receptor model-based Virtual Ligand Screening for the structure-based discovery of selective small molecule antagonists of the androgen receptor [22] and the thyroid hormone receptor [23], as well as antagonists [24] and agonists [25] of the retinoic acid receptor RAR α . In the present paper we provide comprehensive benchmarking of the approach using a large library of the conformational ensembles of NR ligand binding pockets with determined three-dimensional structures. We demonstrate the exceptional performance of the ensemble docking approach on the tasks of predicting the ligand poses, test the ability of the ensembles to recognize their ligands in virtual screening, and propose a method to evaluate ligand selectivity profiles.

Methods

Preparation of receptor structures

Co-crystal structures of ligand binding domains (LBD) of the progesterone receptor (PR), the mineralocorticoid receptor (MR), vitamin D receptor (VDR), the androgen receptor (AR), the peroxisome proliferator activated receptor alpha, beta, gamma (PPAR α , β , γ), the estrogen receptor alpha, beta (ER α , β), the retinoic acid receptor gamma (RAR γ), the glucocorticoid receptor (GR), the thyroid receptor (TR β), the retinoid X receptor alpha (RXR α) were used for this study. PDB entries with X-ray resolution >3 Å were excluded. When several crystal structures were available for a specific receptor/compound complex, the better resolution entry was taken.

PDB files were converted into the internal coordinate representation using the Internal Coordinate Mechanics (ICM) method [26]. In the process of conversion, hydrogen atoms and missing heavy atom were built; polar hydrogens were sampled to optimize the hydrogen-bonding networks. Low occupancy side chains were optimized and assigned the lowest energy conformation. Orientations of ambiguously placed amino acids such as histidine, glutamine and asparagine were optimized to improve the hydrogen bonding patterns. For all NRs except ER only active (agonist-bound) conformations were tested in this study. In case MR and AR, wild type ligand binding domains were studied along with therapeutically relevant point mutants.

Generation of nuclear receptor ligand database

The database included 157 known ligands including agonists, antagonists, and selective nuclear receptor modulators. The molecules were taken from recent publications. For each receptor screened, between three and twenty-three known active ligands were included in the database. Most ligands targeted one of the library receptors, but several ligands had annotated activity against several receptors. Compounds were converted into 3D representation by the ICM conversion procedure. Bond order, tautomeric form, stereochemistry, hydrogen atoms and protonation state were assigned automatically. Each ligand was assigned the MMFF [27] force field atom types and charges and was then subject to Cartesian minimization. The free molecules were optimized with global energy in the internal coordinate space by the ICM method [26, 27]. This procedure guaranteed that the docking procedure did not have any memory of the bound conformation of the ligands.

Internal coordinates mechanics grid docking

Ligand docking simulations were carried out using ICM molecular modeling software (Molsoft LLC, La Jolla, CA)

[26, 27] based on the Biased Probability Monte Carlo (BPMC) simulations. ICM docking procedure performs optimization of the ligand internal coordinates in the set of grid potential maps of the receptors [28]. The ligand binding pocket of each receptor was represented by pre-calculated potential grid maps with grid step of 0.5 Å. The potentials including three van der Waals grid potential (for a carbon probe, large atom probe, or hydrogen probe), a hydrogen bonding grid potential, an electrostatic grid potential, and a hydrophobic grid potential. Ligands were sampled in the obtained maps. The sampling time ranged from 20 s to 3 min per ligand per processor. All calculations were carried out on Intel Core 2 Duo 2.40 GHz.

In the single receptor docking mode, ligands were docked independently into each receptor conformation and score according to ICM full-atom scoring scheme [29]. In the ensemble mode, docking poses of each ligand in all ensemble structures were merged in a single hit list, and the ligand was represented by the top-scoring pose in that list.

ICM full atom scoring

The ICM scoring [28, 29] function was previously derived as a compromise between Gibbs binding energy and numerical errors. It is calculated as follows:

$$E_{\text{bind}} = \Delta E_{\text{IntFF}} + T\Delta S_{\text{Tor}} + \alpha_1 \Delta E_{\text{HBond}} + \alpha_2 \Delta E_{\text{HBDsol}} + \alpha_3 \Delta E_{\text{SolEl}} + \alpha_4 \Delta E_{\text{HPhob}} + \alpha_5 Q_{\text{Size}}$$

where α_i are ligand- and receptor independent constants. ΔE_{IntFF} includes the van der Waals interaction of the ligand with the receptor as well as internal force-field energy of the ligand. $T\Delta S_{\text{Tor}}$ is the change in free energy due to the conformational entropy loss for the ligand upon binding. ΔE_{HBond} , ΔE_{SolEl} , ΔE_{HBDsol} , ΔE_{HPhob} and Q_{Size} are the hydrogen bonding, the solvation electrostatic energy, the hydrogen bond donor/acceptor desolvation correction, the hydrophobic terms, respectively, and Q_{Size} is the number of ligand non-hydrogen atoms.

Virtual ligand screening

The ICM virtual ligand screening module was used. The ligand database was docked into each protein structure. The top scoring pose per ligand was refined and rescored. The accuracy of VLS was tested as the area under the ROC curves obtained plotting the number of false positives vs. the number of true positives in the hit list.

Determination of NRs specific binding energy offsets

For each nuclear receptor NR_k in the benchmark ($k = 1, 2, \dots, 17$), the initial binding score offset was determined as the

negative 5% percentile threshold of the database compounds binding scores. In other words, all compounds in the database were scored and rank-ordered the ensembles of NR_k , producing the set of scores $S(\text{Cmpd}_i, \text{NR}_k)$ ($i = 1, 2, \dots, 157$, $k = 1, 2, \dots, 17$). The threshold T_k was defined so that 5% of the database compounds had scores below this threshold. The new scores for all compounds were calculated as $S_{\text{off}}(\text{Cmpd}_i, \text{NR}_k) = S(\text{Cmpd}_i, \text{NR}_k) - T_k$.

The optimized NR-specific offsets were obtained by the Nelder-Mead simplex [30] optimization of the normalized ROC AUC on the matrix of scores $S_{\text{off}}(\text{Cmpd}_i, \text{NR}_k)$ as a function of the vector of offsets (T_1, T_2, \dots, T_{17}). The available experimental data on compounds activity was used to define true positives, false positives, true negatives, and false negatives. The initial offsets were scattered by addition of random vector of numbers in the ranges of $[-10, 10]$ and used as multiple starting point for optimization. The obtained optimal combination of offsets provided discrimination of binders from non-binders with ROC AUC as high as 90%.

Results

Building a library of structurally distinct high-quality conformational ensembles of nuclear receptor pockets

We collected a library of high-quality crystallographic structures of ligand binding domains of various human nuclear receptors to be used in virtual screening and selectivity profiling of NR ligands. Each receptor was represented by an ensemble of structures rather than a single structure in order to capture possible conformational variations of the binding pocket. For four receptors, estrogen receptors $\text{ER}\alpha$ and $\text{ER}\beta$, mineralocorticoid receptor MR, and androgen receptor AR, we built more than one ensemble.

In case of the estrogen receptors $\text{ER}\alpha$ and $\text{ER}\beta$, separate ensembles represented active and inactive conformations. Depending on the nature of the bound ligand (agonist or antagonist), the carboxy-terminal helix H12 was found in one of two orientations. In the agonist-bound conformation, it serves as a “lid” to close the ligand-binding pocket, whereas in the antagonist-bound conformation, this helix is positioned in a different orientation opening the entrance to the LBP. The classification was performed based on the conformational analysis rather than on the nature of the crystallized ligand. For example, even though genistein, is an agonist of $\text{ER}\beta$, the co-crystal structure (PDB ID: 1qkm [31]) was classified as inactive because of the observed position of H12.

In case of mineralocorticoid receptor MR, all wild-type structures were merged into one ensemble, while all

structures carrying a clinically relevant S810L point mutation formed the second ensemble. This mutation is known to cause severe hypertension, increase the stability of ligand-MR complexes, and change the mode of action of existing antagonists converting them into agonists. For example, progesterone and spironolactone (Fig. 1), antagonists of the wild-type mineralocorticoid receptor [32, 33], act as agonists of S810L MR [34]. Similar situation was observed for the androgen receptor AR, as T877A point mutation allows antagonists of the wild-type receptors to invoke the agonist-bound conformation of the ligand binding domain [35, 36].

For other receptors, the identified point mutations did not change the mode of binding and action of the existing ligands. By this reason, the mutant structures were included in the same ensemble as the wild-type structures.

The structures were further analyzed and filtered to retain only high-resolution (<2.9 Å) X-ray structures that reproduced the binding geometry of their cognate ligands in docking. All protein structures in the ensemble were superimposed using all atoms in the immediate vicinity of the ligands. The superimposition algorithm was based on an iterative procedure that, through an unbiased weight assignment to different atomic subsets, gradually found the better superimposable core of atom pairs between the template and the other structures. The energy-based pre-processing and refinement of the structures was performed that included selection of optimal His tautomers, prediction of orientation of Asn, Gln and His residues, and optimization of polar hydrogen atoms in order to reproduce correct hydrogen bonding networks. The resulting library of NR structure ensembles is summarized in Table 1. It

includes 17 ensembles of 13 nuclear receptors, with three to six structures per ensemble (65 structures total).

Prediction of ligand binding geometry by single receptor cross-docking

We first studied the ability of the collected pocket structures to correctly reproduce the binding geometry of their cognate ligands and other ligands in their ensembles. The flexible compound docking and scoring was performed using the Internal Coordinate Mechanics (ICM) algorithm described in Methods. As the criterion of correct binding pose prediction, the standard ligand heavy-atom RMSD cutoff of 2 Å was used. As we expected, all ligands were reproduced correctly within 2 Å RMSD in the self-docking (Fig. 2 and Table 2).

In cross docking, 63 compounds were docked into the structures of their target receptors, with the cognate (co-crystallized) structure excluded. That made the total of 256 structure-ligand pairs. As shown in Fig. 2, the near-native ligand conformation scored at the top in 202 cases (78%), and was found within three top-scoring poses in 220 cases (85%). The success rate of ligand binding pose prediction for individual receptors is shown in the Table 2.

Most importantly, the predicted ligand poses correctly captured most or all essential ligand-receptor contacts predicted by crystallography, mutagenesis, and quantum mechanical studies [37–39]. This makes the predicted poses good starting points for compound optimization or understanding effects of pocket residue modifications.

Improved prediction of ligand binding geometry by ensemble docking

In multiple receptor conformation, or *ensemble* docking, alternative docking poses of a single ligand from different receptor structures were combined in a single hit list, and the ligand was represented by the best scoring pose in this list. The cognate receptor structure was excluded from the ensemble. As shown in Fig. 2, ensemble docking with excluded cognate receptor structure reproduced the near-native ligand binding geometry as the top-scoring pose in 89% of the cases (56 out of 63 receptor-ligand pairs). 36 of 63 best scoring poses of each ligand in its receptor ensemble are shown in Fig. 3 (the rest of poses can be found in supporting information Figure S2). For three more cases (PPAR δ /GNI, PPAR γ /EHA and ER α /E4D), the correct geometry was found within three top scoring poses in the hit list. The result of this section confirms that using experimentally determined structure ensembles with small number of conformers improves docking accuracy as compared to single receptor mode (89%

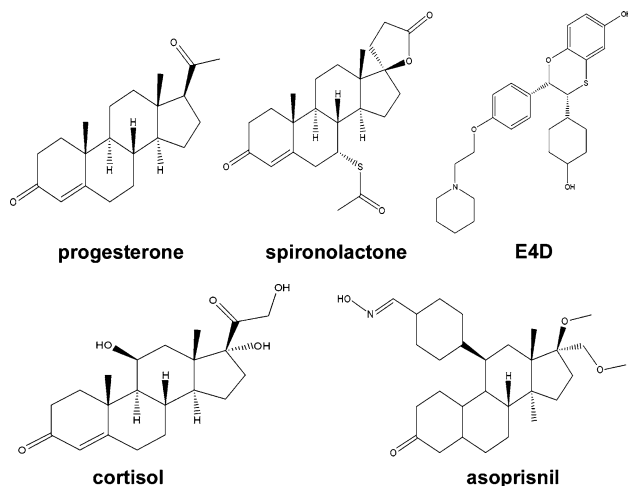


Fig. 1 Chemical structures of compounds mentioned in the text

Table 1 Composition of the nuclear receptor library

	Ensemble ID	Description	PDB IDs and co-crystallized ligand Het IDs
1	ER α –	Estrogen receptor α , inactive conformation	1xp6 (AIU), 1xqc_a (AEJ), 1sj0 (E4D), 2ouz (C3D)
2	ER α +	Estrogen receptor α , active conformation	2b1v_a (458), 2b1z_a (17M), 2fai_a (459), 1l2i_a (ETC.)
3	ER β –	Estrogen receptor β , inactive conformation	1hj1 (AOE), 1qkm (GEN), 1qkn (RAL)
4	ER β +	Estrogen receptor β , active conformation	1u3r_a (338), 1yy4_a (4NA), 1zaf_b (789)
5	GR+	Glucocorticoid receptor, agonist—bound	1m2z_a (DEX), 1p93_a (DEX), 3bqd (DAY)
6	wtMR+	Mineralocorticoid receptor, agonist—bound	2aa7 (1CA), 2a3i (C0R), 2aa2 (AS4),
7	S810L MR+	Mineralocorticoid receptor, agonist—bound	1y9r_a (1CA), 2aa6_a (STR), 1ya3_a (STR), 2aax_a (STR)
8	PR+	Progesterone receptor, agonist—bound	1e3k_a(R18), 1zuc_b (T98), 2ovh (AS0)
9	RAR γ +	Retinoic acid receptor γ , agonist—bound	1exa (394), 1fcx (184), 2lbd (REA)
10	RXR α +	Retinoic acid receptor α , agonist—bound	1fm9 (REA), 1mzn_a (BM6), 1rdt (L79), 1xvp_a (CID), 2p1u (4TN)
11	VDR+	Vitamin D3 receptor, a agonist—bound	1db1 (VDX), 1s19 (MC9), 1txi (TX5), 2har (OCC), 2has (C3O), 2hb7 (O1C)
12	wtAR+	Androgen receptor, agonist—bound	1e3g (R18), 2am9 (TES), 1t7t (DHT), 2hvc (LGD)
13	T877A AR+	Androgen receptor, agonist—bound	1gs4 (ZK5), 2ax7 (FHM), 2oz7 (CA4)
14	PPAR α +	Peroxisome proliferator-activated receptor α , agonist—bound	2npa_c (MMB), 2rew (REW), 2p54 (735), 1k7l_a (544),
15	PPAR δ +	Peroxisome proliferator-activated receptor δ , agonist—bound	1gwx_a (433), 1y0s_a (331), 2j14_b (GN1)
16	PPAR γ +	Peroxisome proliferator-activated receptor γ , agonist—bound	1rdt (L79), 1k74 (544), 1knu (YPA), 1nyx (DRF), 2f4b (EHA), 2gtk (208)
17	TR β +	Thyroid hormone receptor β , agonist—bound	1n46_b (PFA), 1nax (IH5), 1q4x (G24), 2h6w (T3)

(+): agonist-bound or active state, (–): antagonist-bound or inactive state

success rate in ensemble docking vs. 78% in single receptor docking).

Successful recognition of active compounds by the NR pocket ensembles in virtual screening

We tested the ability of the NR pockets and their conformational ensembles to select the active ligands of their respective receptors from a large and challenging compound database. Docking and scoring the database compounds in a single pocket or a pocket ensemble produced an ordered hit list. The screening selectivity of the models is represented by their ability to bring the active compounds to the top of the list, and is evaluated numerically as the Area Under normalized Receiver Operating Characteristic (ROC) curve, or AUC.

The ligand database was collected as described in Methods. It consisted of 157 various nuclear receptor ligands described in literature. 124 of these compounds had annotated agonist or antagonist activity against at least one of the library receptors (supplementary information Table 1). For each receptor, between three and twenty-three known ligands were included in the database. Most compounds in the database targeted exactly one receptor,

but for some of them the activity against two or more receptors was described in literature.

In the *single pocket* mode, 29 out of 66 NR structures (43%) showed high selectivity towards their respective ligands with ROC-AUC exceeding 90%. For 53 structures (80%) ROC-AUC exceeded 80%. The results of VLS with single pocket conformation are summarized in Table 3.

When used in the *ensemble* mode, the ligand hit lists from different structures in the ensemble were combined and each ligand was represented by its best score in the ensemble. This remarkably improved the screening performance of our NR structure library. For 11 of 17 ensembles, ROC-AUC values of more than 90% were achieved, and for others they exceeded 80%. The most significant improvement was obtained for ER α –, ER β – and PR+ which showed poor selectivity in the single receptor mode, but achieved the ROC AUC values of 86, 87 and 92% respectively, in the ensemble mode (Fig. 4). This result suggests that ensemble screening is the most efficient approach to VLS.

Variations in the screening selectivity of the individual ensembles can be in part explained by the incompleteness of our ligand benchmark set. Because in the absence of information on binding of a particular ligand to a receptor of interest we assumed that the ligand is inactive, low

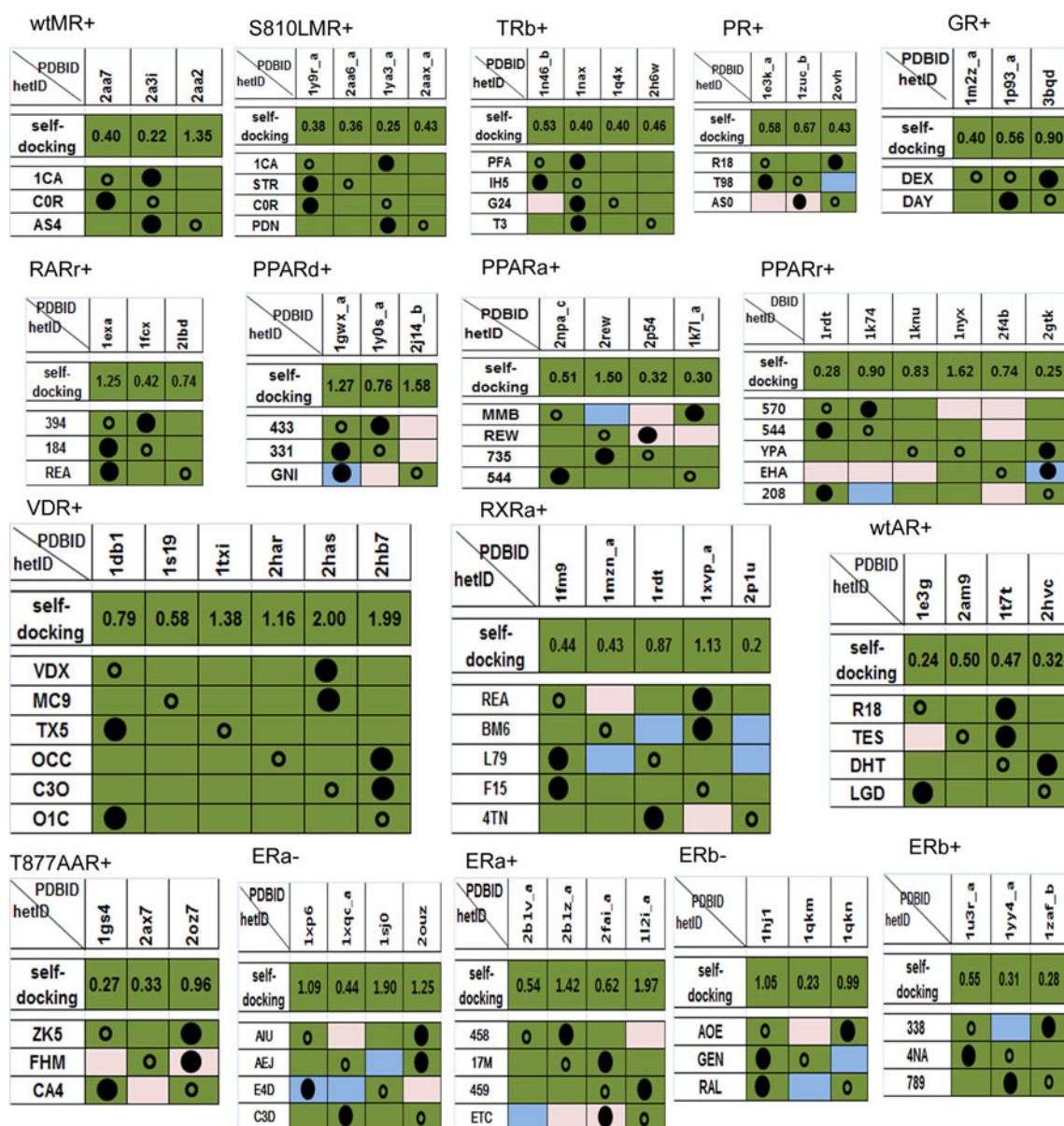


Fig. 2 Self- and cross-docking of crystallographic compounds into individual pockets of the proposed NR pocket library. The table entries are colored by Rmsd of the compound docking poses from the crystallographic poses: *green*—correct pose is ranked first; *blue*—correct pose is found within top three; *pink*—correct pose is not found

within top three poses. *Empty circle*—cognate ligand (hence self-docking); *filled circle*—top score in the ensemble. All compound chemical structures can be found in Supplementary information Figure S1

docking score for that ligand was automatically classified as a false positive. Further examination of several false positives confirmed their activity against the target thus qualifying them as true positives instead. That was the case, for example, for diethylstilbestrol for which ER α agonist activity was confirmed after docking by literature search [40]. Additionally, a number of false positives in the benchmark were due to the fact that antagonist (inactive) binding pockets are often sterically compatible with the

agonist molecules (but not vice versa) and share significant subsets of interactions. The use of pharmacophore-like atomic property fields in the spirit of [41, 42] may help to distinguish and eliminate this ambiguity.

False negatives, i.e. active compounds with poor docking scores may also affect the evaluation of the overall selectivity of the ensembles. Our structural ensembles are able to capture some, but not all aspects of the receptor equilibrium, especially for flexible pockets like ER α . In

Table 2 Successful self- and cross-docking of crystallographic compounds into individual pockets of the proposed NR pocket library

Receptor	Self-docking RMSD	Fraction of correctly predicted structure/ligand pairs in single-receptor cross-docking	
		Self-docking included (%)	Self-docking excluded (%)
wtMR+	0.65±0.60	100	100
S810LMR+	0.35±0.07	100	100
TRβ+	0.44±0.06	93	91
PR+	0.56±0.12	55	33
GR+	0.62±0.25	100	100
RARγ+	0.80±0.41	100	100
PPARδ+	1.20±0.41	55	33
PPARα+	0.65±0.56	75	75
PPARγ+	0.77±0.50	73	62
VDR+	1.31±0.60	100	100
RXRα+	0.61±0.27	76	70
wtAR+	0.38±0.12	93	91
T877AAR+	0.52±0.38	66	50
ERα−	1.17±0.59	68	58
ERα+	1.13±0.59	75	66
ERβ−	0.75±0.45	66	50
ERβ+	0.38±0.14	88	83

RMSD of self-docking is described by average across all pockets in the ensemble ± standard deviation

such cases, addition of new experimental conformations as they become available may help improve the quality of the compound recognition.

Determination of protein-dependent binding free energy offsets for compound selectivity profiling

Ligand selectivity profiling is a very challenging task of computational structural biology that requires an accurate prediction of relative binding energy estimate of a single compound to different receptors. It is *orthogonal* to the task of screening, which is based on prediction of binding energies of a single receptor to multiple database compounds. It has been previously shown that intrinsic properties of the proteins, such as protein entropy loss upon ligand binding, structure quality fluctuations, and inevitable errors in energy functions can be, to a certain level, absorbed into a single protein-dependent binding energy offset which can be pre-calculated using experimentally observed binding energies or statistical considerations [42]. Such offsets must be applied in compound selectivity profiling.

There are several ways to determine the offsets. The easiest, naïve way is based on the distribution of calculated binding scores of all database compounds in a given

ensemble, and does not depend on any experimental data about active compounds. A single offset is calculated for each ensemble and applied in such a way that exactly 5% of all database compounds have binding scores better than a single fixed threshold. A better, more precise way to determine the offset is based on analysis of existing experimental compound activity data. The offsets (one number for each ensemble) are obtained by a machine learning algorithm that optimizes the recognition of correct targets for all database compounds.

In this work, we derived the protein-specific binding energy offsets in both ways, applied them to the predicted ligand binding energies, and evaluated the ligand specificity profiling accuracy of the obtained models. Figure 5 presents the comparison of the naïve offsets with those derived by simplex minimization of the total number of errors in the receptor-ligand binding matrix. It is clear that the naïve offsets are sufficiently close to the trained offsets ($R^2 = 0.64$) and, therefore, represent a reasonable approximation of the true offsets to be used in compound profiling.

Ligand selectivity profiling using the structure library and the precalculated binding energy offsets

We found that even raw ICM docking scores were predictive of compound selectivity profiles. This is demonstrated by Fig. 6a on which the ROC curve for recognition of experimentally confirmed high-affinity ligand receptor pairs in the set of all $157 \times 17 = 2,669$ possible pairs is shown in grey. This type of analysis relies on the ability of the model to deliver comparable values of the docking scores of a single ligand to multiple receptors. Complementing the ligand docking scores with the pre-determined NR-specific binding energy offsets further improved ligand selectivity prediction. The histogram of per-ligand AUC values with and without the offset correction is shown in Fig. 6b. The AUC of 100 for a selected compound is an evidence of *perfect* compound selectivity profile recognition, i.e. the situation in all compound/target-receptor pairs are predicted with better scores than compound/non-target-receptor pairs. As shown in Fig. 6b, the perfect profile recognition was achieved for 53 of 124 compounds without the offsets, and for 66/124 compounds with the offsets. Figure 7 presents a good agreement of the activities predicted by docking with annotated experimental compounds activities. For example, cortisol (Ligand No 46 in the Fig. 8) is known as to be an agonist of GR and T877A AR, which does not bind to wtAR. This profile is correctly predicted by our ensemble docking. Another example is asoprisnil (Ligand No 55), agonist of PR, which does not bind to ER and MR. Biological activity of asoprisnil was well characterized by predicting binding energy in our

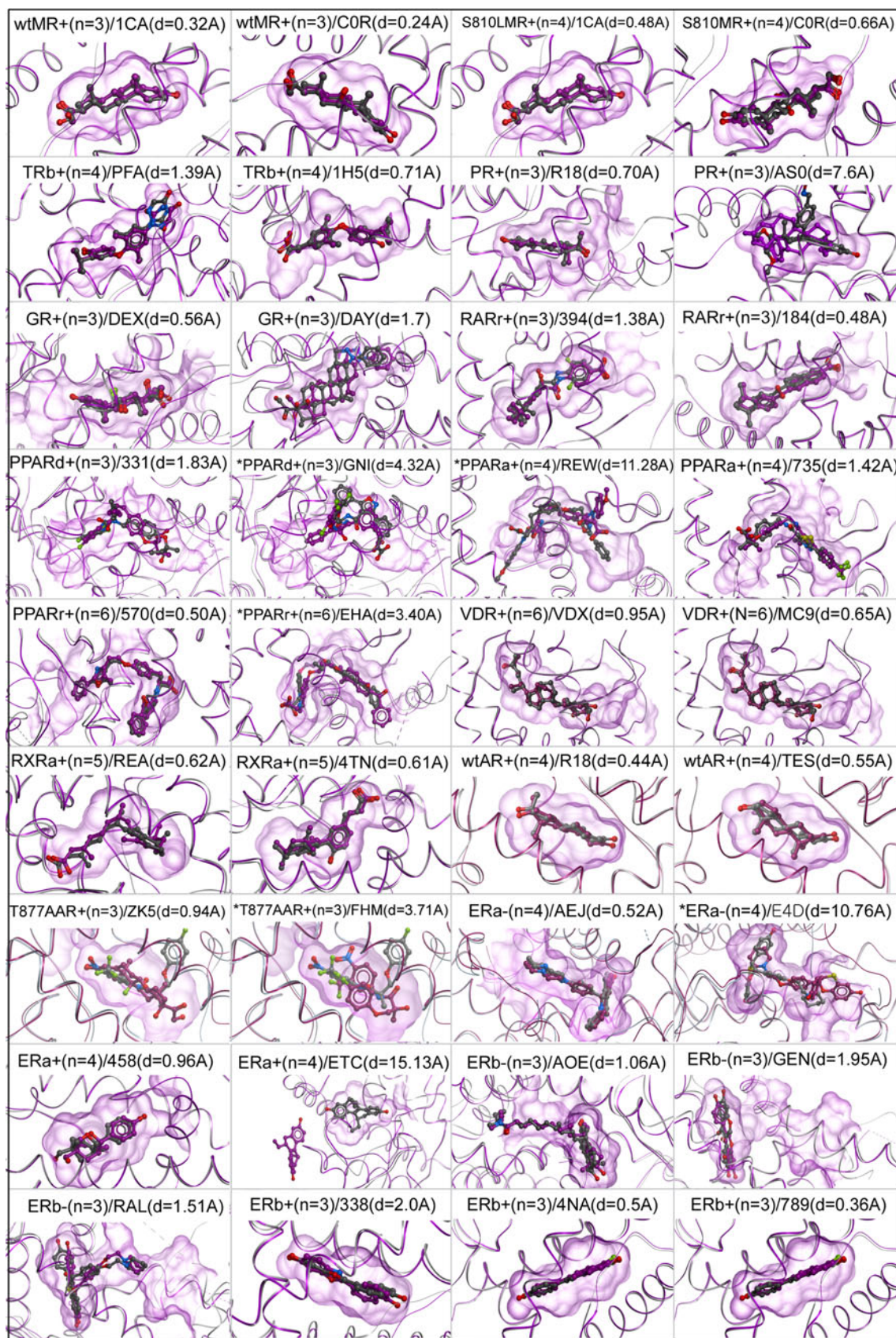


Fig. 3 Comparison of the top-scoring docking pose of the crystallographic ligands in the pocket ensembles of their target NR (magenta) with their crystallographic poses (grey). *The correct geometry was found within three top-scoring poses. For each pose, n is number structures in the ensemble and d is the Rmsd between the native and docking poses

MRC model. The patterns observed in Fig. 7 may also reveal potentially complex cross-reactivity profiles of the NR ligands and provide explanation for the off-target activities or side effects of the studies ligands. Experimental validation of these predictions would allow further refinement of the method.

The obtained results confirm the applicability of the NT binding pocket ensembles in the task of compound selectivity profiling.

Discussion

Proteins are flexible machines that exist in a dynamic conformational equilibrium in solution. They tend to change their conformation, often significantly, to

accommodate binding of ligands (the so-called induced fit effect). Multiple receptor conformation (MRC) docking [20], with its faster 4D incarnation [21], represent one possible approach to incorporation of receptor flexibility in ligand docking. With the rapid growth of structural information in the form of X-ray crystal structures and NMR, it becomes possible to design multi-conformational pocket libraries that provide the unique possibilities for structure-based drug design.

Nuclear receptors form a major class of drug targets. They are involved in a variety of human diseases including cancer, hormonal disorders, and diabetes. In order to understand the structural basis of selectivity of NR-targeting existing drugs, and to develop novel selective nuclear receptor modulators, it is important to build a comprehensive panel of 3D pockets with their conformational variations. In this work we present the first such panel consisting of 17 ensembles of binding pockets of 13 different receptors. We tested ability of the designed ensembles to correctly reproduce the compounds binding poses, draw a distinction between the active compounds and other NR ligands in virtual screening, and predict compound selectivity profiles.

Table 3 Accuracy of VLS in single receptor mode measured by their ROC AUC values

Receptor	PDB ID	ROC %	Receptor	PDB ID	ROC %	Receptor	PDB ID	ROC %
ER α -	1xp6	71	PR+	1zuc_b	60	PPAR α +	2npa_c	88
	1xqc_c	82		1e3k_a	75		2p54	78
	1sj0	73		2ovh	88		2rew	83
	2ouz	89		1exa	97		1k7l_a	92
ER α +	2b1v_a	85	RAR γ +	1fcx	96	PPAR δ +	1y0s_a	95
	2fai_a	80		2lbd	95		1gwx_a	97
	1l2i_a	81		2p1u	90		2j14_b	99
	2b1z_a	80		1rdt	92		1rdt	89
ER β -	1hj1	90	RXR α +	1mzn	88	PPAR γ +	1nyx	86
	1qkn	75		1fm9	86		1k74	89
	1qkm	61		1xvp	92		1knu	90
ER β +	1u3r_a	96		1db1	96		2f4b	87
	1yy4_a	98	VDR+	1txi	78	TR β +	2gtk	91
	1zaf_b	97		2has	97		2h6w	93
GR+	1m2z_a	80		1s19	76		1n46_b	93
	1p93_a	75		2har	76		1q4x	93
	3bdq	65		2hb7	92		1nax	91
wtMR+	2aa7	98	wtAR+	2hvc	81			
	2a3i	96		1e3g	81			
	2aa2	99		1t7t	87			
S810LMR+	1y9r_a	80		2am9	83			
	1ya3_a	95	T877AAR+	1gs4	89			
	2aa6_a	91		2oz7	87			
	2aax_a	93		2ax7	87			

The crystal structure 1rdt is a heterodimer of PPAR γ and RXR α and was used separately for both receptors

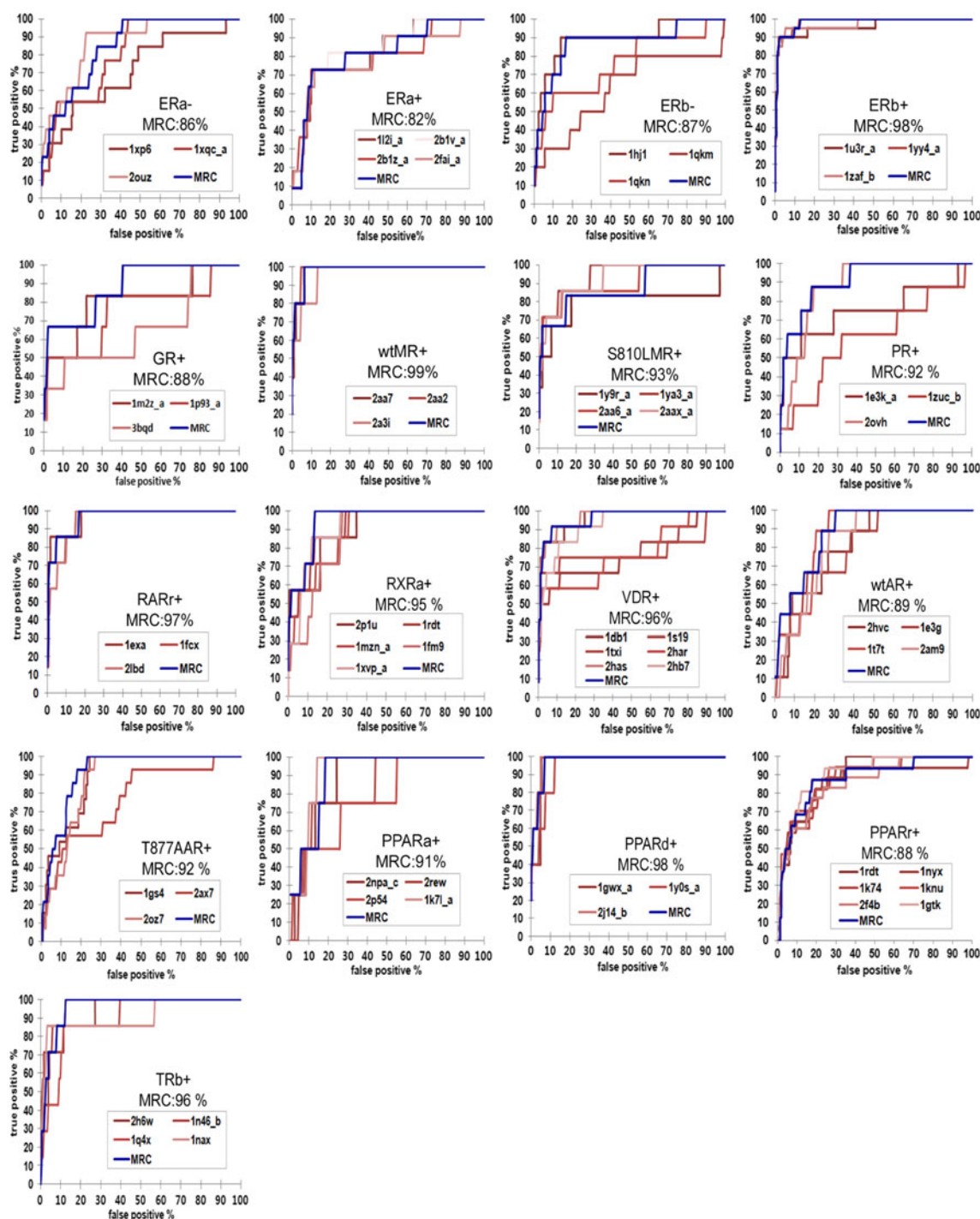


Fig. 4 Comparison of VLS accuracy in single and multiple receptor modes, as measured by the areas under the ROC curves. ROC curves of single and multiple receptor docking are shown in red and blue, respectively. +; agonist, −; antagonist, MRC; Multiple Receptor Conformation

The collected NR pocket ensembles reproduced the near-native ligand binding geometry as the top-scoring pose in 78% of the cases in the single receptor docking mode, and in 89% of the cases in the ensemble mode. The high success rate ensemble docking suggests that the obtained models may be good starting points for

optimization of the existing compounds or virtual ligand screening.

Most NR models demonstrated high screening selectivity even in single rigid receptor mode. The selectivity was further improved by combining them in the ensemble mode. The high selectivity of all 17 ensembles towards

Fig. 5 Comparison of receptor-specific binding energy offsets derived by naïve method and by machine learning

Receptor	initial offset	optimized offset
ER α -	-6.9	-5.7
ER α +	-12.8	-9.2
ER β -	-5.1	-1.3
ER β +	-7.9	-12.9
GR+	0.1	6.4
S810LMR+	2.0	1.7
wtMR+	4.8	7.2
PR+	0.7	5.8
RAR γ +	7.0	-1.1
RXR α +	8.3	9.0
VDR+	1.1	-3.4
wtAR+	-3.3	-3.1
T877AAR+	-6.7	-4.9
PPAR α +	-0.2	1.4
PPAR δ +	8.0	4.2
PPAR γ +	2.4	0.1
TR+	8.4	5.7

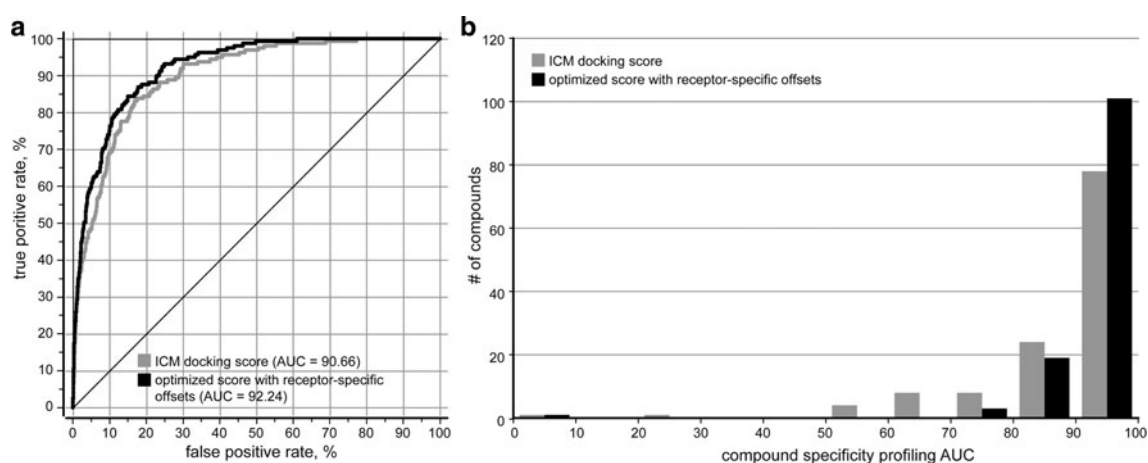
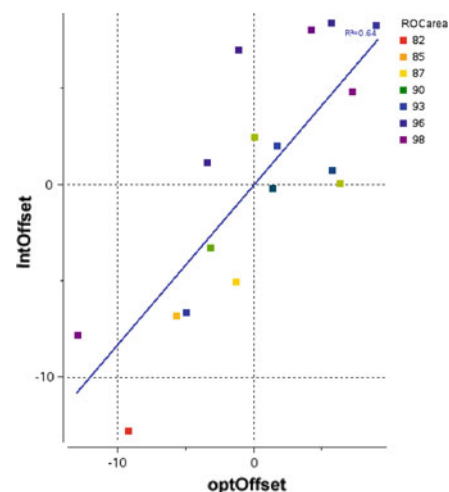


Fig. 6 Recognition of correct receptor-ligand binding pairs and ligand profiling using the proposed NR pocket ensemble library: **a** overall ROC curve with and without receptor-specific offsets;

b histogram of per-compound AUC values. AUC value of 100 indicates perfect correspondence between the predicted and the experimentally determined compound activity

their active ligand is exemplified by the ROC-AUC values above 80%; for 11 ensembles, it exceeded 90%. All receptors recognized most of their active ligands in top 2–10% of the hit list.

Selectivity between different isoforms of the same receptor can represent an important parameter of understanding the binding mode of agonists and antagonists. Here, we built pocket ensembles of PPAR α , δ , γ and ER α , β , and evaluated their ability to discriminate between isotype-specific ligands in virtual screening. All three pocket ensembles of PPAR (α , δ , and γ), recognized their known agonist and antagonists with high selectivity (91, 98 and 88% respectively). Ensembles of ER α , β also showed high

selectivity even though the structures of their ligand binding domains are very similar. Further on, the conformation-specific ensembles of ER α , and ER β were also used to test the selectivity between different conformational states of the same protein. The active and inactive ensembles successfully distinguished agonists from antagonists and vice versa. The obtained results suggest that our proposed pocket ensemble library can potentially be used as a powerful device for structure-based discovery of novel specific nuclear receptor modulators with the desired mechanism of action.

Furthermore, by combining the compound docking scores with pre-determined receptor-specific binding

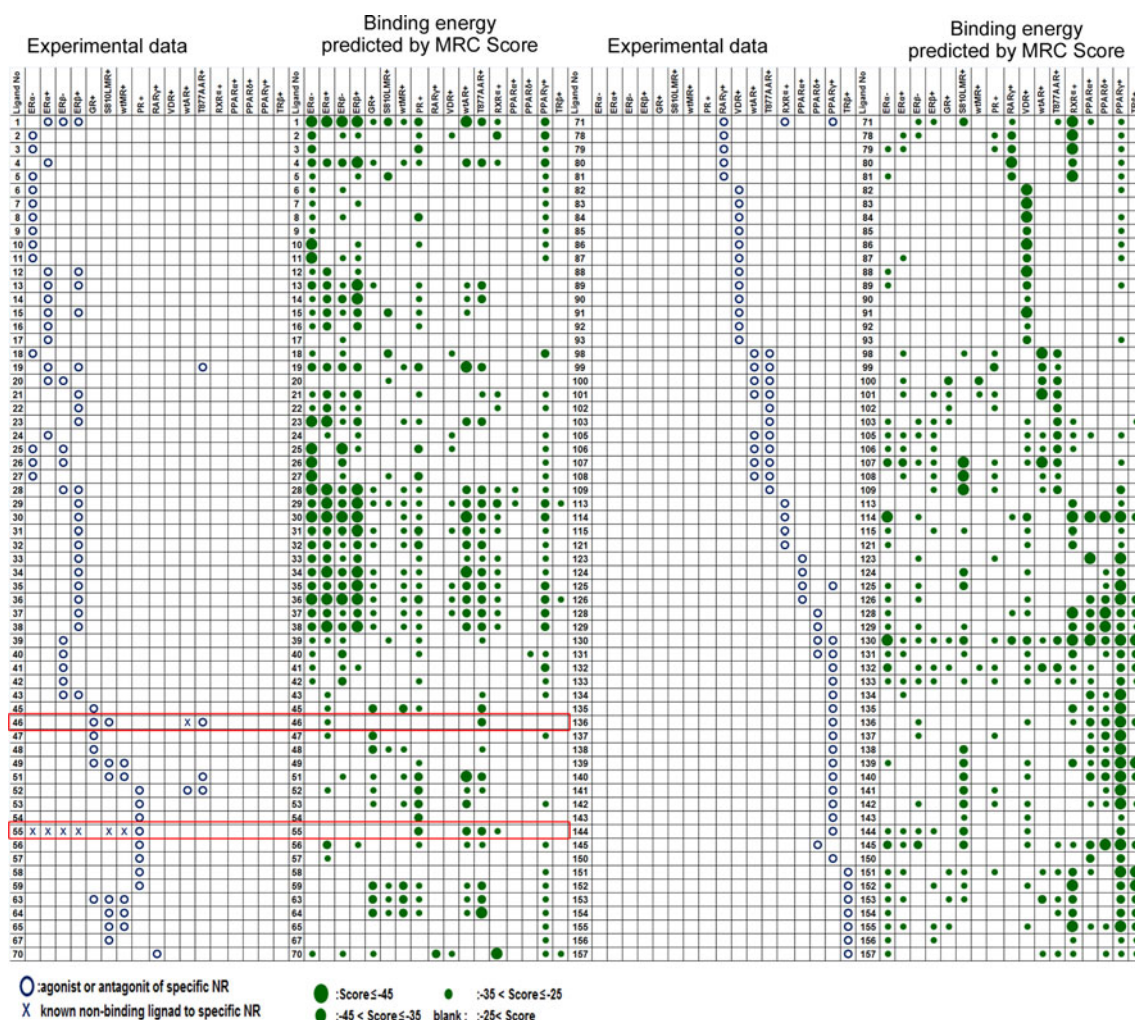


Fig. 7 Comparison of experimental activities of known NR ligands with their binding energy predicted by the proposed pocket ensemble library. In the experimental data panels, *Empty circle* indicates annotated agonist/antagonist activity, *x* indicates *validated* absence of activity. In the predicted binding energy panels, the *green circles* of

energy offsets, we showed that our proposed pocket ensembles can be used for *in silico* determination of compound selectivity profiles. This aspect is extremely important as it can open avenues for rational prediction of possible off-target effects of the experimental compounds, or for repurposing of the existing drugs [22, 43, 44].

Conclusion

In this work, we presented a comprehensive library of three dimensional models of the nuclear receptor binding pockets in different functional states. Using the library, we validate and demonstrate the advantage of using the ensemble docking approach in various structure-based drug discovery applications, including compounds pose prediction, virtual screening, and selectivity profiling.

different size indicate the binding energy predicted by combining the docking score with pre-determined receptor-specific offsets. Ligand No is the same order as the benchmark library in the supplementary information Table 1

Acknowledgments This work was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2007-357- E00038 38) and NIH/NIGMS grants 5-R01-GM071872 and 1-R01 -GM074832.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Chawla A, Repa JJ, Evans RM, Mangelsdorf DJ (2001) Science 294:1866
2. Kastner P, Mark M, Chambon P (1995) Cell 83:859
3. Weatherman RV, Fletcher RJ, Scanlan TS (1999) Annu Rev Biochem 68:559
4. Mahajan MA, Samuels HH (2005) Endocr Rev 26:583

5. Bourguet W, Germain P, Gronemeyer H (2000) *Trends Pharmacol Sci* 21:381
6. Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schutz G, Umesono K, Blumberg B, Kastner P, Mark M, Chambon P, Evans RM (1995) *Cell* 83:835
7. Bonta PI, Pols TW, de Vries CJ (2007) *Trends Cardiovasc Med* 17:105
8. Bachmann K, Patel H, Batayneh Z, Slama J, White D, Posey J, Ekins S, Gold D, Sambucetti L (2004) *Pharmacol Res* 50:237
9. Pascucci JM, Gerbal-Chaloin S, Duret C, Daujat-Chavanieu M, Vilarem MJ, Maurel P (2008) *Annu Rev Pharmacol Toxicol* 48:1
10. Dees EC, Kennedy MJ (1998) *Curr Opin Oncol* 10:517
11. Kolvenbag GJ, Iversen P, Newling DW (2001) *Urology* 58:16
12. Lydon JP, DeMayo FJ, Funk CR, Mani SK, Hughes AR, Montgomery CA Jr, Shyamala G, Conneely OM, O'Malley BW (1995) *Genes Dev* 9:2266
13. Pitt B, Zannad F, Remme WJ, Cody R, Castaigne A, Perez A, Palensky J, Wittes J (1999) *N Engl J Med* 341:709
14. Bouillon R, Okamura WH, Norman AW (1995) *Endocr Rev* 16:200
15. Ingraham HA, Redinbo MR (2005) *Curr Opin Struct Biol* 15:708
16. Nettles KW, Greene GL (2005) *Annu Rev Physiol* 67:309
17. Rochel N, Moras D (2006) *Curr Top Med Chem* 6:1229
18. Togashi M, Borngraeber S, Sandler B, Fletterick RJ, Webb P, Baxter JD (2005) *J Steroid Biochem Mol Biol* 93:127
19. Sperandio O, Miteva MA, Delfaud F, Villoutreix BO (2006) *Curr Protein Pept Sci* 7:369
20. Totrov M, Abagyan R (2008) *Curr Opin Struct Biol* 18:178
21. Bottegoni G, Kufareva I, Totrov M, Abagyan R (2009) *J Med Chem* 52:397
22. Bisson WH, Cheltsov AV, Bruey-Sedano N, Lin B, Chen J, Goldberger N, May LT, Christopoulos A, Dalton JT, Sexton PM, Zhang XK, Abagyan R (2007) *Proc Natl Acad Sci USA* 104:11927
23. Schapira M, Raaka BM, Das S, Fan L, Totrov M, Zhou Z, Wilson SR, Abagyan R, Samuels HH (2003) *Proc Natl Acad Sci USA* 100:7354
24. Schapira M, Raaka BM, Samuels HH, Abagyan R (2000) *Proc Natl Acad Sci USA* 97:1008
25. Schapira M, Raaka BM, Samuels HH, Abagyan R (2001) *BMC Struct Biol* 1:1
26. Abagyan R, Totrov M (1994) *J Mol Biol* 235:983
27. Abagyan R, Totrov M, Kuznetsov DA (1994) *J Comp Chem* 15:488
28. Totrov M, Abagyan R (1997) *Proteins Suppl* 1:215
29. Schapira M, Totrov M, Abagyan R (1999) *J Mol Recognit* 12:177
30. Nelder JA, Mead R (1965) *The Computer Journal* 7:308
31. Pike AC, Brzozowski AM, Hubbard RE, Bonn T, Thorsell AG, Engstrom O, Ljunggren J, Gustafsson JA, Carlquist M (1999) *EMBO J* 18:4608
32. Rogerson FM, Yao YZ, Smith BJ, Dimopoulos N, Fuller PJ (2003) *J Mol Endocrinol* 31:573
33. Quinkler M, Johanssen S, Bumke-Vogt C, Oelkers W, Bähr V, Diederich S (2001) *Mol Cell Endocrinol* 171:21
34. Geller DS, Farhi A, Pinkerton N, Fradley M, Moritz M, Spitzer A, Meinke G, Tsai FT, Sigler PB, Lifton RP (2000) *Science* 289:119
35. Bohl CE, Gao W, Miller DD, Bell CE, Dalton JT (2005) *Proc Natl Acad Sci USA* 102:6201
36. Sun C, Shi Y, Xu LL, Nageswararao C, Davis LD, Segawa T, Dobi A, McLeod DG, Srivastava S (2006) *Oncogene* 25:3905
37. Ito M, Fukuzawa K, Ishikawa T, Mochizuki Y, Nakano T, Tanaka S (2008) *J Phys Chem B* 112:12081
38. Ito M, Fukuzawa K, Mochizuki Y, Nakano T, Tanaka S (2007) *J Phys Chem A* 112:1986
39. Fukuzawa K, Mochizuki Y, Tanaka S, Kitaura K, Nakano T (2006) *J Phys Chem B* 110:16102
40. Miller CP, Collini MD, Tran BD, Harris HA, Kharode YP, Marzolf JT, Moran RA, Henderson RA, Bender RHW, Unwalla RJ, Greenberger LM, Yardley JP, Abou-Gharbia MA, Lyttle CR, Komm BS (2001) *J Med Chem* 44:1654
41. Totrov M (2008) *Chem Biol Drug Des* 71:15
42. Kufareva I, Abagyan R (2008) *J Med Chem* 51:7921
43. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijter MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KLH, Edwards DD, Shoichet BK, Roth BL (2009) *Nature* 462:175
44. Fan H, Irwin JJ, Webb BM, Klebe G, Shoichet BK, Sali A (2009) *J Chem Inf Model* 49:2512